ERIC Notebook

March / April 2001 Issue 17

# Common Measures and Statistics in Epidemiological Literature.

Michel Ibrahim, MD, PhD
Director of Education Program

Lorraine Alexander, DrPH

Carl Shy, MD, DrPH

Sandra Deming, MPH
Graduate Research Asst.

www.sph.unc.edu/courses/eric

Ron Horner, PhD
ERIC Director

Beth Armstrong
ERIC Program Manager

http://hsrd.durham.med.va.gov/ERIC/

The ERIC Notebook is funded by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP), to promote the strategic growth of the epidemiologic capacity of the DVA.

For the non-epidemiologist or non-statistician, understanding the statistical nomenclature presented in journal articles can sometimes be challenging, particularly since multiple terms are often used interchangeably, and still others are presented without definition. This notebook will provide a basic introduction to the terminology commonly found in epidemiological literature

## Measures of disease frequency

Measures of disease frequency characterize the occurrence of disease, or death in a population. These measures are descriptive in nature and indicate how likely one is to develop a disease in a specified population. The three most common measures of disease frequency are incidence, incidence rate, and prevalence.

**Incidence (Risk):**

*Incidence*, also known as *risk*, *cumulative incidence*, *incidence proportion*, or *attack rate* (Although not rate at all) is a measure of the probability of a unaffected individual developing a specified disease over a given period of time. *Risk* usually refers to the individual while *incidence proportion* refers to the population under study. For a given period of time (i.e.: 1 month, 5 years, lifetime):

$$\text{Incidence} = \frac{\#\text{ of new cases of a given disease}}{\text{total \# individuals at risk}}$$

A 5-year incidence or risk of .10 indicates that an individual at risk has a 10% chance of developing the given disease over a 5-year period of time.

Risk is generally measured in prospective studies as the population at risk can be defined at the start of the study, and followed for the development of disease. However, risk cannot be measured directly in case-controls studies as the total population at risk cannot be defined. Thus, in case-control studies, a group of diseased and non-diseased individuals are selected, and the odds of developing the disease, is calculated as opposed to calculating risk.

$$\text{odds} = \frac{\#\text{ individuals with the disease}}{\#\text{ individuals without the disease}}$$

**Incidence Rate:**

*Incidence rate* is a measure of how quickly disease or death is occurring in a population. The numerator is the same as in incidence, but the denominator includes a measure of time, typically person-years. (Person-time is defined as the sum of time that each at-risk individual contributes to the study.

$$\text{Incidence rate} = \frac{\#\text{ of new cases of a given disease}}{\text{Sum of the person-time at risk}}$$

Thus an incidence rate of .1 case/person-years indicates that, on average, for every 10 person-years (i.e.: 10 people each followed 1 year or 2 people followed for 5 years, etc.) contributed, 1 new case of disease will develop.

**Prevalence:**

*Prevalence* is the proportion of a population who has the disease at a given period of time. *Prevalence* is generally the preferred measure when it is difficult to define onset of disease (such as asthma), or any disease of long duration (chronic conditions such as arthritis). A limitation of the prevalence measure is that it tends to favor the inclusion of chronic diseases over acute ones. Also, inferring causality is troublesome with prevalence data, as typically both the exposure and outcome are measured at the same time. Thus it may be difficult to determine if the suspected cause precedes the outcome of interest.

Prevalence = $\frac{\text{\# of affected individuals}}{\text{total \# individuals in the population}}$

Thus a population with a heart disease prevalence of 0.25 indicates that 25% of the population is affected by heart disease at a specified moment in time. A final note, incidence and incidence rates can also refer to the incidence of death in a population and are termed *mortality* and *mortality rate respectively*.

## Measures of association

*Measures of association* are utilized to compare two or more populations, typically those with differing exposure or disease status, to identify factors with possible etiological roles in disease onset. Note that evidence of an association does not imply that the relationship is causal; the association may be artifactual or non-causal as well. Common measures of association include the *risk difference, relative risk (also known as risk ratio)*, and the *odds ratio.*

**Risk Difference (Risk Ratio):**

*Risk difference* is defined as

$= \text{Risk}_{\text{exposed}} - \text{Risk}_{\text{unexposed}}$:

$\frac{\text{\# cases in exposed group}}{\text{total \# at risk in exp. group}} - \frac{\text{\# cases in control group}}{\text{total \# at risk control grp.}}$

The risk difference, also know as the attributable risk, provides the difference in risk between two groups indicating how much excess risk is due to the exposure of interest. A positive risk difference indicates excess risk due to the exposure, while a negative result indicate that the exposure of interest has a protective effect against the outcome. (Vaccinations would be a good example of an exposure with a protective effect). This measure if often utilized to determine how much risk can be prevented by an effective intervention.

**Relative Risk:**

The *relative risk* (RR) is commonly found in cohort studies and is defined as: the ratio of the risk in the exposed group to the risk in the unexposed group.

$$\text{Risk (incidence)}_{\text{exposed}} \;/\; \text{Risk (incidence)}_{\text{unexposed}} = \text{RR}$$

The relative risk is a measure of the strength of the association between the exposure and the outcome. How is the relative risk interpreted? A RR of 1.0 indicates there is no difference in risk between the exposed and unexposed group. A RR greater than 1.0 indicates a positive association, or increased risk for developing the disease in the exposed group. A relative risk of 1.50 indicates that the exposed group has a 50% or a two-fold increase in risk of having the outcome as compared to the unexposed group. Inversely, a relative risk of less than 1.0 indicates a negative association between the exposure and outcome in the exposed group compared to the unexposed group. In this case, the exposure provides a protective effect. For example, a relative risk of 0.80 where the exposed group received a vaccination indicates that the risk of disease is 25% lower in the exposed group as compared to the unexposed group.

One of the benefits the measure risk difference has over the risk ratio is that it provides the absolute difference in risk, information that is not provided by the ratio of the two. A relative risk of 2.0 can imply both a doubling of a very small or large risk, and one cannot determine which is the case unless the individual risks are presented.

**Odds Ratio:**

The third measure of association is the *odds ratio* (OR). The formula for the OR is:

$$\text{odds ratio} = \frac{\text{odds}_{\text{exposed}}}{\text{odds}_{\text{unexposed}}}$$

It is determined in place of the relative risk in case-control studies. In this type of study, the underlying population at risk for developing the disease cannot be determined because individuals are selected as either diseased or non-diseased. An odds ratio can approximate the relative risk in instances where the disease prevalence is low (Less that 10%), otherwise there is a tendency for the OR to overestimate the relative risk.

The odds ratio is interpreted in the same manner as the relative risk with and OR of 1.0 indicating no association, an OR greater than 1.0 indicating a positive association, and an OR less than 1.0 indicating a negative, or protective association.

**The Null Value:**

The null value is a number corresponding to there being no effect, that is, no association between exposure and disease. In epidemiology, the null value for relative risk is 1.0, and it is also 1.0 for odds ratios and prevalence ratios (terms you will use in later modules). A relative risk of 1.0 is obtained when the risk of disease among exposed is equal to the risk of disease among the nonexposed. Statistical testing

focuses on the null hypothesis, which is a statement predicting that there will be no association between exposure and disease (or between the assumed cause and its effect), i.e. that the relative risk will equal 1.0. If the data obtained from a study provide evidence against the null hypothesis, then this hypothesis can be rejected and an alternative hypothesis becomes highly probable. For example, a null hypothesis would say that there is no association between children having cigarette smoking mothers and the incidence of asthma in those children. If a study showed that there was a greater incidence of asthma among such children (compared with children of nonsmoking mothers), and that the relative risk of asthma among children of smoking mothers was 2.5 with a 95% confidence interval of 1.7 to 4.0, we would reject the null hypothesis. The alternative hypothesis could be expressed in two ways: 1) children of smoking mothers will have either a higher or lower incidence of asthma than other children, or 2) children of smoking mothers will only have a higher incidence of asthma. The first alternative hypothesis involves what is called a "two-tailed test" and is used when we simply have no basis for predicting in which direction from the null value exposure is likely to be associated with disease, or, in other words, whether exposure is likely to be beneficial or harmful. The second alternative hypothesis involves a "one-tailed test" and is used when we have a reasonable basis to assume that exposure will only be harmful (or if we were studying a therapeutic agent, that it would only be beneficial).

## Measures of significance

**The P-Value:**

The "p" value is an expression of the probability that the difference between the observed value and the null value has occurred by "chance", or more precisely, has occurred simply because of sampling variability. The smaller the "p" value, the less likely the probability that sampling variability accounts for the difference. Typically, a "p" value less than 0.05, is used as the decision point, meaning that there is less than a 5% probability that the difference between the observed relative risk and 1.0 is due to sampling variability. If the "p" value is less than 0.05, the observed relative risk is said to be "statistically significant." The exclusive use of "p" values for interpreting results of epidemiologic studies has been strongly discouraged in the more recent texts and literature because research on human health is not conducted to reach a decision point ( a "go" or "no go" decision), but rather to obtain evidence that there is reason for concern about certain exposures or lifestyle practices or other factors that may adversely influence the health of the public. Statistical tests of significance, using "p" values, were developed for industrial quality-control purposes, in order to make a decision whether the manufacture of some item is achieving acceptable quality. We are not making such decisions when we interpret the results of research on human health.

The lower bound of the 95% confidence interval is also often utilized to decide whether a point estimate is statistically significant, i.e. whether the measure of effect (e.g. the ratio 2.5 with a lower bound of 1.8) is statistically different than the null value of 1.0.

## Measures of precision

**Confidence Interval:**

A *confidence interval* expresses the extent of potential variation in a point estimate (the mean value or relative risk). This variation is attributable to the fact that our point estimate of the mean or relative risk is based on some sample of the population rather than on the entire population. For example, from a clinical trial, we might conclude that a new treatment for high blood pressure is 2.5 times as effective as the standard treatment, with a 95% confidence interval of 1.8 to 3.5. 2.5 is the point estimate we obtain from this clinical trial. But not all subjects with high blood pressure can be included in any study, thus the estimate of effectiveness, 2.5, is based on a particular sample of people with high blood pressure. If we assume that we could draw other samples of persons from the same underlying population as the one from which subjects were obtained for this study, we would obtain a set of point estimates, not all of which would be exactly 2.5. Some samples would be likely to show an effectiveness less than 2.5, and some greater than 2.5. The 95% confidence interval tells us that we are 95% confident that these studies will yield a point estimate in the range of 1.8 to 3.5. Thus we can also say that the new treatment for high blood pressure is 2.5 times as effective as the standard treatment, but this measure could range, with 95% confidence, from a low of 1.8 to a high of 3.5.

The confidence interval also provides information about how precise an estimate is. The tighter, or narrower, the confidence interval, the more precise the estimate. Typically, larger sample sizes will provide a more precise estimate. Estimates with wide confidence intervals should be interpreted with caution.

## Other Terms

**Crude and Adjusted Values:**

There are often two types of estimates presented in research articles, *crude* and *adjusted* values. *Crude* estimates refer to simple measures that do not

account for other factors that may be driving the estimate. For instance, a crude death rate would simply be the number of deaths in a calendar year divided by the average population for that year. This may be an appropriate measure in certain circumstances, but could become problematic if you want to compare two or more populations that vary on specific factors known to contribute to the death rate. For example, you may want to compare the death rate for two populations, one of which, is located in a high air pollution area, to determine if air pollution levels affect the death rate. The high air pollution population may have a higher death rate, but you also determine that it is a much older population. As older individuals are more likely to die, age may be driving the death rate, rather than the pollution level. To account for the difference in age distribution of the populations, one would want to calculate an *adjusted* death rate that adjusts for the age structure of the two groups. This would remove the effect of age from the effect of air pollution on mortality.

Adjusted estimates are a means of controlling for confounders or accounting for effect modifiers in analyses. Some factors that are commonly adjusted for include gender, race, socioeconomic status, smoking status, and family history.

## Self-Evaluations

**Q1: Select the correct answer:**

a. The odds ratio and relative risk should always approximate one another.
b. The odds ratio will tend to overestimate the relative risk unless the disease prevalence is less than 10%.
c. The relative risk will tend to overestimate the odds ratio unless the disease prevalence is less than 10%.
d. Odds ratios are calculated in cohort studies and relative risks are calculated in case-control studies.

**Q2: Based on the table, calculate the requested measures.**

| | **Diseased** | **Non-diseased** | **Total** |
|---|---|---|---|
| **Exposed** | **450** | **300** | **750** |
| **Non-Exposed** | **100** | **500** | **600** |
| **Total** | **650** | **800** | **1300** |

a. Determine the odds ratio comparing the exposed to the non-exposed.

b. Calculate the relative risk comparing the exposed to the non-exposed.

c. Calculate the risk difference between the exposed and non-exposed groups.

d. Assuming the disease in this situation is a chronic disease, calculate its prevalence in the entire sampled population.

**Q3: Which of these effect estimates would you be most confident in concluding that an association does exist, and why?**

a. OR = 0.95 (95% CI = 0.70 - 1.29)

b. OR = 9.85 (95% CI = 0.89 - 36.48)

c. OR = 1.89 (95% CI = 1.81 - 1.95)

d. OR = 1.91 (95% CI = 1.02 - 2.89)

**Answers to Self Evaluation:**

**Q1:** The correct answer is b. The odds ratio will tend to overestimate the relative risk unless the disease prevalence is low. This is the case because the odds ratio does not account for the underlying risk in the population since only diseased and non-diseased individuals are sampled rather than the population.

**Q2:**

**a.** Odds ratio = $\frac{\text{odds exposed}}{\text{odds unexposed}}$

$= \frac{450 / 500}{100 / 300} = 7.50$

The odds of having the disease in 7.5 times greater in the exposed group as compared to the unexposed group.

**b.** Relative Risk = $\frac{\text{Risk}_{\text{exposed}}}{\text{Risk}_{\text{unexposed}}}$

$= \frac{450 / 750}{100 / 600} = 3.60$

A relative risk of 3.60 indicates that the risk in the exposed group in 3.6 times greater than the risk in the unexposed group.

**c.** The risk difference in the = $Risk_{exposed} - Risk_{unexposed}$:

450 / 750 - 100 / 600 = .4333 excess cases in the exposed group for every one case in the unexposed group.

.4333 is the excess disease risk in the exposed group compared to the unexposed group. To interpret this value, for every 10 cases of disease occurring in the unexposed group, there are 14 cases in the exposed group.

**d.** Prevalence = $\frac{\text{\#cases at a given point in time}}{\text{total \# at risk at that time}}$

Prevalence = $\frac{650}{1300} = .50$

At the time when this data was collected, the prevalence of the disease in the sampled population was 50%.

**Q3:**

The best answer is c. The odds ratio indicates a moderate association, and the confidence interval, that does not include 1.0, indicates that the effect estimate is precise and significant. Answer a. has a confidence interval that includes that null value. You should be cautious about making conclusion about answer b. because although there is a very strong association suggested with an OR= 9.85, the estimate is very imprecise as demonstrated by the wide confidence interval. This imprecision may be due to a small study population. Answer d. also suggests that a moderate association exists, however the confidence interval is larger than that in answer c. suggesting that the estimate is less precise.

**Reminder:**

ERIC Notebook can be viewed online at

http://sph.unc.edu/courses/eric or

http://hsrd.durham.med.va.gov/ERIC/Education/Education.html.

**ERIC Notebooks for CME Credit !!!!**

Did you know that now you can earn CME credit for reading and answering a few questions about the ERIC notebook?

This option is currently available on-line for this notebook, with future notebooks to follow. To find out how to earn CME credit, go to:

http://cdlhc.sph.unc.edu/courses/eric/

**ERIC NOTEBOOK IS PRODUCED BY THE EDUCATIONAL ARM (MICHEL A. IBRAHIM, MD, PHD, DIRECTOR) OF THE EPIDEMIOLOGIC RESEARCH AND INFORMATION CENTER AT DURHAM, NC (RON HORNER, PHD, DIRECTOR)**

**If you are not currently receiving ERIC Notebook and would like to be added to the distribution list, please fill out the form below:**

Name: _____

Degree(s):_____

Address:_____

City, State, Zip:_____

Telephone Number:_____

Fax Number:_____

E-mail Address:_____

Place of Employment: VA____ Other Fed. Govt____
Academic____ Private _____

**Please fax to**: 919-416-5836 - Attn: Beth Armstrong **Or**

**Mail to** : Beth Armstrong, ERIC Program Manager, VA Medical Center (152), 508 Fulton Street, Durham, NC 27705

## Upcoming Topics

* **Common Statistical Tests in Epidemiological Literature**
* **Causality**
* **Health Care Epidemiology**

**Please let Beth Armstrong know which topics are of special interest to you so that we can include them in a future issue.**

**COMING SOON !**

**In the near future, you will be able to earn CME credit for reading and answering questions about the ERIC notebooks.**

**Check upcoming issues for details**


BETH ARMSTRONG, ERIC PROGRAM MANAGER
VA MEDICAL CENTER (152)
508 FULTON STREET
DURHAM, NC 27705